DOCUMENT RESUME

ED 338 629 TM 013 058

AUTHOR Marsh, Herbert W.
TITLE Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research.
PUB DATE 87
NOTE 30p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Students; *Construct Validity; Factor Analysis; Feedback; Foreign Countries; Higher Education; *Professors; *Research Methodology; *Student Evaluation of Teacher Performance; *Teacher Effectiveness; Test Construction; Test Reliability; Test Validity
IDENTIFIERS Australia; Multidimensional Models; Papua New Guinea; Spain; *Students Evaluation of Educational Quality

ABSTRACT

H. W. Marsh's monograph (1987) on students' evaluations of teaching effectiveness in higher education is summarized. The research, which emphasized the construct validity approach, led to the development of the Students' Evaluations of Educational Quality (SEEQ) instrument. Factor analysis resulted in identification of nine SEEQ factors--learning value, instructor enthusiasm, organization, individual rapport, group interaction, breadth of coverage, examinations and grading, assignments and readings, and workload difficulty. The analysis encompassed 5,000 classes conducted for five groups of courses selected to represent diverse academic disciplines at the graduate and undergraduate levels. Instructors evaluated their own teaching effectiveness on the same SEEQ form as that completed by their students. Tertiary students from different countries evaluated teaching effectiveness with the SEEQ. Use of the instrument indicates that class-average student ratings are: (1) multidimensional; (2) reliable and stable; (3) primarily a function of the instructor rather than of course content; (4) relatively valid against a variety of indicators of effective teaching; (5) relatively unaffected by a variety of variables hypothesized as potential biases; and (6) perceived as useful feedback by faculty about their teaching, by students for use in course selection, and by administrators for use in personnel decisions. Eight data tables and one flowchart are presented. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Herbert W. Marsh
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED338629

Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions For Future Research

[A Summary of the monograph by Marsh, H.W. (1987), International Journal of Educational Research, 11, 253-387 (Whole Issue).]

Herbert W. Marsh

The University of Sydney, Australia

ABSTRACT

The purpose of this presentation is to summarize my research on students' evaluations of teaching effectiveness in higher education. The research led to the development of the Students' Evaluations of Educational Quality (SEEQ) instrument. These findings indicate that class-average student ratings are:

1) multidimensional;

2) reliable and stable;

3) primarily a function of the instructor who teaches a course rather than of the course that is taught;

4) relatively valid against a variety of indicators of effective teaching;

5) relatively unaffected by a variety of variables hypothesized as potential biases; and

6) seen to be useful by faculty as feedback about their teaching, by students for use in course selection, and by administrators for use in personnel decisions.

TM013058

BEST COPY AVAILABLE

Chapter 1: Introduction.

Historical Perspective

Students have evaluated teachers for as long as there have been individuals claiming to be teachers. Programs of formal collection of students' evaluations were introduced in the United States in the 1920s. H. H. Remmers iniated an extensive research program in the 1920's that spanned 3 decades and anticipated many of the issues presently considered. The topic has been one of the most frequently studied and controversial in American educational research.

Purposes For Collecting Students' Evaluations

Students' evaluations of teaching effectiveness are commonly collected at most North American universities. Appropriate purposes of these evaluations are to provide:

1) diagnostic FEEDBACK to faculty about the effectiveness of their teaching;

2) a measure of teaching effectiveness to be used in PERSONNEL DECISIONS;

3) information for students to use in INSTRUCTOR/COURSE SELECTION;

4) an outcome or a process description for RESEARCH ON TEACHING;

It will be argued here that students' evaluations as typically defined are not appropriate for the evaluation of courses -- as opposed to the instructors who teach the courses.

Construct Validity Approach

My research emphasizes a construct validity approach to the study of students' evaluations of teaching and several perspectives that underlie this approach:

** effective teaching and students' evaluations designed to reflect it are multidimensional/multifaceted;

** there is no single criterion of effective teaching; and

** tentative interpretations of relations with validity criteria and potential biases must be scrutinized in different contexts and must examine multiple criteria of effective teaching.

Chapter 2: The Dimensionality Students' Evaluations

Student ratings and the teaching that they represent are MULTIDIMENSIONAL (e.g., a teacher may be quite well organized but lack enthusiasm).

Information from students' evaluations depends upon the content of the items. Poorly worded or inappropriate items will not provide useful information. If a survey instrument contains an ill-defined hodge-podge of different items and student ratings are summarized by an average of these items, then there is no basis for knowing what is being measured.

Surveys should contain separate groups of related items which are:

1) supported by empirical procedures such as factor analysis;

2) derived from a logical analysis of the content of effective teaching and the purposes which the ratings are to serve, or a carefully constructed theory;

Factor Analysis.

Empirical techniques such as factor analysis provide a test of whether:

1) students differentiate among different components of effective teaching;

2) the empirical factors match the ones the instrument was designed to measure;

3) there is a large halo effect -- a generalization from some subjective feeling, an external influence or an idiosyncratic response mode -- that affects responses to all the items.

***Factor analysis cannot determine whether the obtained factors are important to understanding effective teaching. This requires a logical analysis of the content of the factors.

Logical Analysis. In the development of SEEQ:

1) a large item pool was obtained from a literature review, forms in current usage, and interviews with faculty and students about what they see as effective teaching;

2) students and faculty were asked to rate the importance of items;

3) faculty were asked to judge the potential usefulness of the items as a basis for feedback;

4) open-ended student comments were examined to determine if important aspects had been excluded.

***These criteria, along with psychometric properties, were used to select items and revise subsequent versions. This systematic development constitutes evidence for the content validity of SEEQ and makes it unlikely that it contains any trivial factors.

The SEEQ Factors (and an example item):

Learning/Value: You have found the course intellectually challenging and stimulating;

Instructor Enthusiasm: Instructor was dynamic and energetic in conducting the course;

Organization: Course materials were well prepared and carefully explained;

Individual Rapport: Instructor was friendly towards individual students;

Group Interaction: Students were encouraged to participate in class discussions;

Breadth of Coverage: Instructor presented background or origin of ideas/concepts developed in class;

Examinations/Grading: Feedback on examinations/graded materials was valuable;

Assignments/Readings: Readings, homework, etc. contributed to appreciation and understanding of subject;

Workload/Difficulty: Course difficulty relative to other classes was (very easy ...medium...very hard)

Factor Analytical Results

Factor analyses identify the factors which SEEQ was designed to measure, and demonstrate that the students' evaluations do measure distinct components of teaching effectiveness.

1) factor analyses of evaluations from 5,000 classes were conducted for 5 groups of courses selected to represent diverse academic disciplines at graduate and undergraduate levels; each clearly identified the SEEQ factors.

2) Instructors were asked to evaluate their own teaching effectiveness on the same SEEQ form as completed by their students. Factor analyses of student ratings and instructor self-evaluations each identified the same SEEQ factors.

3) Tertiary students in different countries (Australia -- University of Sydney; Australia -- TAFE; Papua New Guinea; Spain) evaluated teaching effectiveness with SEEQ. Similar factors were identified for each of the four groups. The items judged to be most important were also similar in these very different educational settings.

***The SEEQ results provide clear support for the multidimensionality of students' evaluations. Students' evaluations cannot be adequately interpreted if this multidimensionality is ignored.

Table 1

Nineteen Instructional Rating Dimensions Adapted From Feldman (1976

---

1) Teacher's stimulation of interest in the course and subject matter.
2) Teacher's enthusiasm for subject or for teaching.
3) Teacher's knowledge of the subject.
4) Teacher's intellectual expansiveness and breadth of coverage.
5) Teacher's preparation and organization of the course.
6) Clarity and understandableness of presentations and explanations.
7) Teacher's elocutionary skills.
8) Teacher's sensitivity to, and concern with, class level and progress.
9) Clarity of course objectives and requirements.
10) Nature and value of the course material including its usefulness and relevance.
11) Nature and usefulness of supplementary materials and teaching aids.
12) Difficulty and workload of the course.
13) Teacher's fairness and impartiality of evaluation of students; quality of exams.
14) Classroom management.
15) Nature, quality and frequency of feedback from teacher to students.
16) Teacher's encouragement of questions and discussion, and openness to the opinions of others.
17) Intellectual challenge and encouragement of independent thought.
18) Teacher's concern and respect for students; friendliness of of the teacher.
19) Teacher's availability and helpfulness.

Note. These nineteen categories were originally presented by Feldman (1976) but in subsequent studies (e.g., Feldman, 1984) "Perceived Outcome or impact of instruction" and "Personal Characteristics ('Personality')" were added while rating dimensions 12 and 14 presented above were not included.

BEST COPY AVAILABLE

Table 2

Factor Analyses of Students' Evaluations of Teaching Effectiveness (S) and the Corresponding Faculty Self-Evaluations of Their Own Teaching (F) in 329 Courses (Reprinted with permission from Marsh, 1984b).

Factor Analyses of Students' Evaluations of Teaching Effectiveness (S) and the Corresponding Faculty Self-Evaluations of Their Own Teaching (F) in 329 Courses

	Factor pattern loadings																	
	1		2		3		4		5		6		7		8		9	
Evaluation items (paraphrased)	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F	S	F
1. Learning/Value																		
Course challenging/stimulating	42	40	23	25	08	−10	04	04	00	−03	15	27	09	05	16	23	29	20
Learned something valuable	53	77	15	02	10	−02	09	04	01	01	10	00	10	04	17	09	16	06
Increased subject interest	57	70	12	06	08	07	08	07	02	−03	18	08	03	−04	19	05	14	−02
Learned/understood subject matter	55	52	12	12	13	12	06	03	03	11	02	−01	19	07	14	−04	−23	−11
Overall course rating	36	33	25	29	16	09	12	08	09	02	12	16	13	−08	14	27	08	16
2. Enthusiasm																		
Enthusiastic about teaching	15	29	55	42	16	00	07	02	21	15	10	00	06	16	01	09	05	06
Dynamic & energetic	08	03	80	70	15	01	11	06	08	06	06	06	07	16	01	05	06	03
Enhanced presentations with humor	10	04	66	58	−04	06	06	01	13	02	12	02	14	07	02	−18	−07	−10
Teaching style held your interest	09	12	59	64	23	20	16	06	06	00	03	14	10	06	06	03	−02	−03
Overall instructor rating	12	27	49	54	23	09	14	08	23	02	11	16	10	−08	05	27	05	16
3. Organization																		
Instructor explanations clear	12	00	07	24	55	42	20	09	06	04	10	06	13	01	06	23	−08	−03
Course materials prepared & clear	06	06	03	−02	73	69	09	01	10	−02	09	04	06	03	10	03	01	12
Objectives stated & pursued	19	12	−06	−08	49	41	03	06	08	06	14	08	25	27	06	05	06	06
Lectures facilitated note taking	−03	02	20	09	58	53	−17	07	−02	06	14	04	15	06	08	01	−04	−06
4. Group Interaction																		
Encouraged class discussions	04	06	10	02	01	03	84	86	03	00	00	00	06	00	06	−05	00	−03
Students shared ideas/knowledge	02	08	06	−07	−04	−01	85	88	06	13	06	01	08	−02	08	−10	−02	01
Encouraged questions & answers	03	−04	06	09	14	06	62	69	16	−02	15	03	07	11	08	21	00	01
Encouraged expression of ideas	07	01	02	06	01	−11	73	75	20	08	06	07	08	12	06	09	00	−02
5. Individual Rapport																		
Friendly towards students	−04	10	17	06	00	−06	13	12	68	78	−01	−06	13	02	10	−06	−07	01
Welcomed seeking help/advice	04	−10	06	02	02	07	06	00	86	75	−04	04	12	06	06	20	03	−04
Interested in individual students	07	10	11	09	00	01	14	07	89	77	−01	−08	14	03	08	−09	03	09
Accessible to individual students	02	−13	−11	−11	16	09	09	−02	62	43	30	25	08	13	00	14	04	07
6. Breadth of Coverage																		
Contrasted implications	−06	02	12	01	06	03	08	01	−03	01	72	84	08	−03	14	02	08	−06
Gave background of ideas/concepts	08	03	08	10	16	07	−03	−02	02	−02	71	78	01	08	11	−01	03	03
Gave different points of view	04	−06	04	08	11	11	08	16	06	01	72	86	07	17	01	−06	04	08
Discussed current developments	23	29	08	−04	−04	−04	06	12	09	00	50	48	08	06	16	10	−01	−02
7. Examinations/Grading																		
Examination feedback valuable	−03	01	08	09	06	−11	09	06	08	12	−04	03	72	62	06	−03	09	03
Eval. methods fair/appropriate	06	02	00	−03	03	14	07	06	14	00	10	17	89	84	11	11	−08	04
Tested emphasized course content	08	00	−01	04	11	21	01	01	06	00	11	−04	79	88	07	10	−02	−03
8. Assignments																		
Reading/texts valuable	−06	08	−03	−03	03	07	−01	−06	03	01	07	−07	01	11	91	70	02	04
Added to course understanding	12	01	−01	−12	01	04	09	21	01	17	−02	08	07	05	81	56	06	10
9. Workload/Difficulty																		
Course difficulty (Easy-Hard)	−06	00	06	−01	04	−06	−04	02	−01	00	08	00	−04	08	10	04	85	74
Course workload (Light-Heavy)	14	−04	−09	−01	03	02	07	06	00	04	06	01	00	01	00	−04	88	86
Course pace (Too Slow-Too Fast)	−20	07	12	00	04	18	−12	−09	06	02	−03	−07	04	08	06	−04	62	32
Hours/week outside of class	14	00	07	00	−11	00	07	02	00	02	−04	03	03	−08	06	21	73	46

Note. Factor loadings in boxes are the loadings for items designed to measure each factor. All loadings are presented without decimal points. Factor analyses of student ratings and instructor self-ratings consisted of a principal-components analysis, Kaiser normalization, and rotation to a direct oblimin criterion. The analyses were performed with the commercially available Statistical Package for the Social Sciences (SPSS) routine (see Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975).

BEST COPY AVAILABLE

Chapter 3: Reliability, Stability and Generalizability

Reliability

The reliability of the class-average response depends upon the number of students rating the class. The reliability of SEEQ factors is about:

1) .95 for 50 students/class

2) .90 for 25 students/class

3) .74 for 10 students/class

4) .60 for 5 students/class

5) .23 for 1 students/class

****Given a sufficient number of students, the reliability of students' evaluations compares favorably with that of the best objective tests.

Long Term Stability

Some critics suggest that students cannot recognize effective teaching until after being called upon to apply course materials in further coursework or after graduation. According to this argument, former students who evaluate courses with the added perspective of time will differ systematically from students who have just completed a course when evaluating teaching effectiveness. However, cross-sectional studies have shown good correlational agreement between the retrospective ratings of former students and those of currently enrolled students.

In a longitudinal study the same students evaluated classes at the end of the course and again several years later, at least one year after graduation. End-of-class ratings in 100 courses correlated .83 with the retrospective ratings (a correlation approaching the reliability of the ratings), and the median rating at each time was nearly the same.

Generalizability -- Teacher and Course Effects

Researchers have also asked how highly correlated student ratings are in two different courses taught by the same instructor, and in the same course taught by different instructors. This research is designed to address two related questions.

1) What is the generality of the construct of effective teaching as measured by students' evaluations of teaching?

2) What is the relative importance of the effect of the instructor who teaches a class on students' evaluations, compared to the effect of the particular class being taught? (If the impact of the particular course is large, then the practice of comparing ratings of different instructors for tenure/promotion decisions may be dubious).

In order to answer these questions I arranged ratings of 1364 courses into sets such that each set contained ratings of:

1) the SAME INSTRUCTOR teaching the SAME COURSE on two occasions (the correlation was .72 for Overall Instructor Rating);

2) the SAME INSTRUCTOR teaching two DIFFERENT COURSES (the correlation was .61);

3) the SAME COURSE taught by a DIFFERENT INSTRUCTOR (the correlation was -.05).

****A more detailed analysis of these results shows that student ratings primarily reflect the effectiveness of the instructor rather than the influence of the course.

Table 3

Long-Term Stability of Student Evaluations: Relative and Absolute Agreement Between End-of-Term and Retrospective Ratings (Reprinted with permission from Overall & Marsh, 1980).

Long-Term Stability of Student Evaluations: Relative and Absolute Agreement Between End-of-Term and Retrospective Ratings

Evaluation items	Correlations between end-of-term and retrospective ratings: Relative agreement — Individual students (N = 1,374)	Correlations between end-of-term and retrospective ratings: Relative agreement — Class-average responses (N = 100)	M differences between end-of-term and retrospective ratings: Absolute agreement — Retrospective ratings (N = 100 classes)	M differences between end-of-term and retrospective ratings: Absolute agreement — End-of-term ratings (N = 100 classes)	M differences between end-of-term and retrospective ratings: Absolute agreement — Difference ratings (N = 100 classes)
1. Purpose of class assignments made clear	.55**	.81**	6.63	6.61	+.02
2. Course objectives adequately outlined	.58**	.84**	6.61	6.47	+.14*
3. Class presentations prepared and organized	.62**	.79**	6.67	6.54	+.13
4. You learned something of value	.53**	.81**	6.65	6.87	−.22**
5. Instructor considerate of your viewpoint	.58**	.83**	6.59	6.88	−.29**
6. Instructor involved you in discussions	.56**	.84**	6.63	6.75	−.12
7. Instructor stimulated your interest	.58**	.82**	6.38	6.50	−.12
8. Overall instructor rating	.65**	.84**	6.55	6.74	−.19*
9. Overall course rating	.56**	.83**	6.65	6.50	+.15*
Median across all nine rating items	.58	.83	6.63	6.61	

Note. A total of 1,374 student responses from 100 different sections each assessed instructional effectiveness at the end of each class (end of term) and again 1 year after graduation (retrospective follow-up). All ratings were made along a 9-point response scale that varied from 1 (very low or never) to 9 (very high or always).

* $p < .05$. ** $p < .01$.

BEST COPY AVAILABLE

Table 4

Correlations Among Different Sets of Classes for Student Ratings and Background Characteristics (Reprinted with permission from Marsh, 1984b).

Correlations Among Different Sets of Classes for Student Ratings and Background Characteristics

Measure	Same teacher, same course	Same teacher, different course	Different teacher, same course	Different teacher, different courses
Student rating				
Learning/Value	.696	.563	.232	.069
Enthusiasm	.734	.613	.011	.028
Organization/Clarity	.676	.540	−.023	−.063
Group Interaction	.699	.540	.291	.224
Individual Rapport	.726	.542	.180	.146
Breadth of Coverage	.727	.481	.117	.067
Examinations/Grading	.633	.512	.066	−.004
Assignments	.681	.428	.332	.112
Workload/Difficulty	.733	.400	.392	.215
Overall course	.712	.591	−.011	−.065
Overall instructor	.719	.607	−.051	−.059
Mean coefficient	.707	.523	.140	.061
Background characteristic				
Prior subject interest	.635	.312	.563	.209
Reason for taking course (percent indicating general interest)	.770	.448	.671	.383
Class average expected grade	.709	.405	.483	.356
Workload/difficulty	.773	.400	.392	.215
Course enrollment	.846	.312	.593	.058
Percent attendance on day evaluations administered	.406	.164	.214	.045
Mean coefficient	.690	.340	.491	.211

BEST COPY AVAILABLE

Chapter 4: VALIDITY

Student ratings, which constitute one measure of teaching effectiveness, are difficult to validate since there is no single criterion of effective teaching.

A construct validation approach requires student ratings to be:

1) substantially correlated with a variety of other indicators of effective teaching; and

2) less correlated with other variables that are not logically related to effective teaching (e.g., potential biases).

Other possible criteria of effective teaching would include :

1) student learning (the most widely accepted);

2) instructor self-evaluations (so long as ratings are not the basis of personnel decisions);

3) evaluations by peers and/or administrators who actually attend class sessions;

4) the frequency of occurrence of specific behaviors observed by trained observers;

5) evaluations of former students at time of graduation or several years later;

Multisection Validity Studies.

Multisection courses are large courses in which students are divided into separate groups (sections) that are independently taught by different instructors according to the same course outline and with the same final examination. The critical question is whether those instructors who receive the best evaluations are the ones whose students perform best on the final examination.

In the ideal multisection validity study:

1) there are many sections;

2) students are randomly assigned to sections or at least enroll without any knowledge about the sections or who will teach them;

3) there are good pretest measures;

4) each section is taught completely by a separate instructor;

5) each section has the same course outline, textbooks, course objectives, and final examination;

6) the final examination is constructed to reflect the common objectives by some person who does not actually teach any of the sections, and, if there is a subjective component, is graded by an external person.

Cohen (1981) conducted a meta-analysis of all known multisection validity studies of students' evaluations. Across 68 multisection courses, student achievement was consistently correlated with student ratings of Skill (0.50), Overall Course (0.47), Structure (0.47), Student Progress (0.47), and Overall Instructor (0.43). Only ratings of Difficulty had a near-zero or a negative correlation with achievement.

Cohen's meta-analysis demonstrates that: sections for which instructors are evaluated more highly by students tend to do better on standardized examinations. This finding supports the validity of the ratings.

Instructor Self-Evaluations.

Instructors' self-evaluations are a good criterion of teaching effectiveness for validating student ratings because:

1) They can be collected in all classes where student ratings are collected;

2) They are likely to be widely accepted as one indicator of effective teaching (so long as personnel decisions are not tied to the responses);

3) Instructors can be asked to evaluate themselves with the same SEEQ instrument used by their students, thereby testing the validity of SEEQ.

In two studies a large number of instructors evaluated their own teaching on essentially the same SEEQ survey which was completed by their students. In both studies:

1) separate factor analyses of teacher and student responses identified the same evaluation factors;

2) student-teacher agreement on every dimension was significant (median rs of 0.49 and 0.45).

3) mean differences between student and faculty responses were small (i.e., student ratings were not systematically higher or lower than faculty self-evaluations).

4) Student/teacher agreement on matching factors (i.e., student ratings of Learning/Value and instructor self-ratings of Learning/Value was high (median rs of 0.49 & 0.45).

5) Student/teacher agreement on nonmatching factors (e.g., student ratings of Organization and instructor self-ratings of Group Interaction) was low (as it should be).

****This means that student-instructor agreement is specific to each factor and cannot be explained in terms of a generalized agreement.

These two studies have important implications:

1) the good student/teacher agreement provides strong support for the validity of student ratings;

2) The specificity of student/teacher agreement to each rating factor supports the multidimensionality of effective teaching.

Ratings By Peers.

Peer ratings, based upon actual classroom visitation, are often proposed as indicators of effective teaching. In studies where peer ratings are NOT based upon classroom visitation, ratings by peers agree with student ratings. However, it is likely that peer ratings are based upon information from students.

Peer ratings that are based upon classroom visitation do not appear to be substantially correlated with student ratings, any other indicator of effective teaching, or even the impressions of other peers. These findings suggest peer evaluations should NOT be used for personnel decisions.

Murray (1980, p. 45), in comparing student ratings and peer ratings, found peer ratings to be:

(1) less sensitive, reliable, and valid;

(2) more threatening and disruptive of faculty morale; and

(3) more affected by non-instructional factors than student ratings.

Summary and Implications of Validity Research.

Student ratings are significantly and consistently related to a number of varied criteria including the ratings of former students, student achievement in multisection validity studies, faculty self-evaluations of their own teaching effectiveness, and, perhaps, the observations of trained observers on specific processes such as teacher clarity. This provides support for the construct validity of the ratings.

Peer ratings, based upon classroom visitation, and research productivity were shown to have little correlation with students' evaluations, and since they are also relatively uncorrelated with other indicators of effective teaching, their validity as measures of effective teaching is problematic.

Table 5

Multitrait-Multimethod Matrix: Correlations Between Student Ratings and Faculty Self-Evaluations in 329 Courses (Reprinted with permission from Marsh, 1984b).

Multitrait-Multimethod Matrix: Correlations Between Student and Faculty Self-Evaluations in 329 Courses

	Instructor self-evaluation factor									Student evaluation factor								
Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Instructor self-evaluations																		
1. Learning/Value	(83)																	
2. Enthusiasm	29	(82)																
3. Organization	12	01	(74)															
4. Group Interaction	01	03	−15	(90)														
5. Individual Rapport	−07	−01	07	02	(82)													
6. Breadth	13	12	13	11	−01	(84)												
7. Examinations	−01	08	26	09	15	20	(76)											
8. Assignments	24	−01	17	05	22	09	22	(70)										
9. Workload/Difficulty	03	−01	12	−09	06	−04	09	21	(70)									
Student evaluations																		
10. Learning/Value	46	10	−01	08	−12	09	−04	08	02	(95)								
11. Enthusiasm	21	54	−04	−01	−02	−01	−03	−09	−09	45	(96)							
12. Organization	17	13	30	−03	04	07	09	00	−05	52	49	(93)						
13. Group Interaction	19	05	−20	52	00	−02	−14	−04	−08	37	30	21	(98)					
14. Individual Rapport	03	03	−05	13	28	−19	−03	−02	00	22	35	33	42	(96)				
15. Breadth	26	15	09	00	−14	42	00	09	02	49	34	56	17	15	(94)			
16. Examinations	18	09	01	−01	06	−09	17	−02	−06	48	42	57	34	50	33	(93)		
17. Assignments	20	03	02	09	−01	04	−01	45	12	52	21	34	30	29	40	42	(92)	
18. Workload/Difficulty	−06	−03	04	00	03	−03	12	22	69	06	02	−05	−05	08	18	−02	20	(87)

Note. Values in parentheses in the diagonals of the upper left and lower right matrices, the two triangular matrices, are reliability (coefficient alpha) coefficients (see Hull & Nie, 1981). The underlined values in the diagonal of the lower left matrix, the square matrix, are convergent validity coefficients that have been corrected for unreliability according to the Spearman Brown equation. The nine uncorrected validity coefficients, starting with Learning, would be .41, .48, .25, .46, .25, .37, .13, .36, & .54. All correlation coefficients are presented without decimal points. Correlations greater than .10 are statistically significant.

BEST COPY AVAILABLE

Chapter 5: Relationship to Background Characteristics: Potential Biases in Students' Evaluations

The construct validity of students' evaluations requires them to be:

1) substantially correlated with indictors of effective teaching (i.e., they are valid):

2) but relatively uncorrelated with variables that are not (i.e., they are not biased).

My research indicates that student ratings are not substantially influenced by potential biases, but that faculty still believe that they are.

In a survey I conducted at the university where SEEQ was developed faculty indicated that student ratings were useful and that teaching quality should be given more emphasis in personnel decisions. Nevertheless they felt that student ratings were biased and other measures of teaching effectiveness are even more biased.

***A dilemma existed in that faculty wanted teaching to be evaluated, but were dubious about any procedure to accomplish this purpose.

How Much Do Potential Biases Affect Students' Evaluations.

In several large studies the combined effect of a large number of potential biases was able to explain a total of between 5% and 20% of the variance in student ratings.

Student ratings were positively correlated with Prior Subject Interest, Expected Grades, and Workload/Difficulty, and specific components of the ratings (e.g., Group Interaction and Individual Rapport) are negatively correlated with class size.

The size of the influence of background characteristics is not huge, but large enough to be worrisome IF THESE RELATIONSHIP REALLY REPRESENT BIASES TO STUDENT RATINGS. However, a more detailed examination of the effects suggests that the relations represent the influence of variables that really do affect teaching effectiveness in a way that is validly reflected in the student ratings.

WORKLOAD/DIFFICULTY EFFECT. Paradoxically, at least based upon the supposition that Workload/Difficulty is a potential "bias" to student ratings, higher levels of Workload/Difficulty were positively correlated with student ratings.

****Since the direction of the Workload/Difficulty effect is opposite to that predicted as a potential bias effect, Workload/difficulty does not appear to constitute a bias to student ratings.

CLASS SIZE EFFECT. Class size is negatively correlated with student ratings of Group Interaction and Individual Rapport but not with other SEEQ factors. Similarly, class size is negatively correlated with instructor self-evaluations for these two factors but not other SEEQ factors.

****The findings argue that class size does have a moderate effect on these two aspects of effective teaching and these effects are accurately reflected in the student ratings.

PRIOR SUBJECT INTEREST EFFECT. The effect of Prior Subject Interest on SEEQ scores was greater than that of any of the 15 other background variables that I considered. For both student ratings and instructor self-evaluations, Prior Subject Interest was most highly correlated with Learning/Value.

****Again the findings suggest that Prior Subject Interest is a variable which influences some aspects of effective teaching (particularly Learning/Value) and these effects are accurately reflected in both the student ratings and instructor self-evaluations. Higher student interest in the subject apparently creates a more favorable learning environment and facilitates effective teaching, and this effect is reflected in student ratings as well as instructor self-evaluations.

Expected Grades. Class-average expected grades are positively correlated with student ratings. There are, however, three quite different explanations for this finding:

1) The "grading leniency hypothesis" proposes that instructors who give higher-than-deserved grades will receive higher-than-deserved student ratings, and represents a serious bias.

2) The "validity hypothesis" proposes that better Expected Grades reflect better student learning, and that a positive correlation between student learning and student ratings supports the validity of student ratings.

3) A "student characteristics hypothesis" proposes that pre-existing student characteristics may affect student learning, student grades, and teaching effectiveness, so that the expected grade effect can be explained in terms of other variables.

While these explanations of the expected grade effect have quite different implications, they are not mutually exclusive. The grade a student receives is likely to be related to the grading leniency of the teacher, how much he/she learned, and characteristics that he/she brought into the course. Not surprisingly there is some support for each explanation.

***It is possible that a grading leniency effect may produce a bias in student ratings, but support for this suggestion is weak and the size of such would be small.

Table 5.2

Path Analysis Model Relating Prior Subject Interest, Reason for Taking Course, Expected Grade and Workload/Difficulty to Student Ratings (Reprinted with permission from Marsh, 1984b)

	Factor											
	I. Prior Subject Interest			II. Reason (General Interest Only)			III. Expected Course Grade			IV. Workload/ Difficulty		
Student ratings	DC	TC	Orig r	DC	TC	Orig r	DC	TC	Orig r	DC	TC	Orig r
Learning/Value	36	44	44	15	13	15	26	20	29	17	17	12
Enthusiasm	17	23	23	09	08	09	20	16	20	11	11	06
Organization	−04	−04	−03	16	16	16	03	02	01	04	04	00
Group Interaction	21	28	29	06	06	07	30	27	31	06	06	−02
Individual Rapport	−05	09	09	−01	−02	−02	18	16	17	06	06	01
Breadth	−07	−03	−03	23	19	19	06	−01	−02	21	21	15
Exams/Grading	−05	03	03	12	10	10	25	18	18	20	20	10
Assignments	11	19	20	21	17	18	19	09	13	30	30	23
Overall course	23	32	33	19	15	16	26	15	22	30	30	23
Overall instructor	12	20	20	13	11	12	24	17	20	17	17	10
Variance components[a]	2.9%	5.1%	5.3%	2.3%	1.5%	1.8%	4.5%	2.6%	4.0%	3.6%	3.6%	1.8%

Note. The methods of calculating the path coefficients (*p* values in Figure 5.1), Direct Causal Coefficients (DC), and total Causal Coefficients (TC) are described by Marsh (1980a). Orig r = original student rating. See Figure 5.1 for the corresponding path model.

[a] Calculated by summing the squared coefficients, dividing by the number of coefficients, and multiplying by 100%.

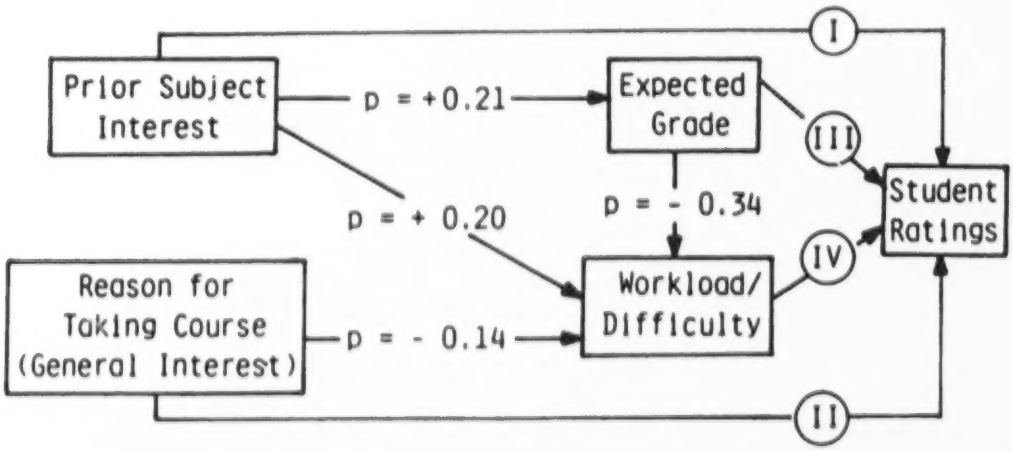

Figure 5.1 Path analysis model relating prior subject interest, reason for taking course, expected grade, and Workload/Difficulty (Path coefficients for the student rating factors appear in Table 5.2; reprinted with permission from Marsh, 1984b)

Chapter 6: 'Dr. Fox' Studies

The Dr. Fox effect is defined as the overriding influence of instructor expressiveness on students' evaluations. In the original Dr. Fox study, a professional actor was favorably evaluated when he lecturered in an enthusiastic/expressive manner, even though he presented material of little educational value. The authors of the study and critics agree that it had serious methodological problems.

To overcome some problems Ware and Williams developed the standard Dr. Fox paradigm in which a series of experimentally manipulated lectures were videotaped. Lectures varied in the content coverage and the expressiveness of delivery. Students viewed one lecture, evaluated teaching effectiveness, and completed an achievement test based on all the material in the high-content lecture. Expressiveness affected student ratings more than did content, whereas content affected achievement test scores more than expressiveness (see meta-analysis by Abrami, et al., 1982).

A Reanalysis.

Marsh and Ware (1982) reanalyzed data from the Ware and Williams studies. A factor analysis of the rating instrument identified five factors which varied in the way they were affected by the manipulations. In the condition most like the university classroom (students knew about the test and a reward prior to viewing the lecture) THE DR. FOX EFFECT WAS NOT FOUND. The instructor expressiveness manipulation only affected rating of Instructor Enthusiasm, the factor most logically related to that manipulation. Content coverage significantly affected ratings of Instructor Knowledge and Organization/Clarity, factors most logically related to that manipulation.

When students had no incentive to perform and did not know they would be tested, instructor expressiveness had a much larger affect on all five student rating factors. In this condition, however, expressiveness also had a larger impact on test scores than the content manipulation. This is one of the few studies to demonstrate that instructor expressiveness causes better examination performance.

How Should the Dr. Fox Effect Be Interpreted?

These results are frequently used to argue for the invalidity of student ratings but my interpretation is quite different. Using a construct validity approach, a specific rating factor should be substantially influenced by manipulations most logically related to it and less influence by other manipulations. This interpretation offers strong support to the validity of student ratings with respect to instructor expressiveness and limited support to their validity with respect to content.

Table 6.1
Effect Sizes of Expressiveness, Content, Expressiveness × Content Interaction in Each of the Three Incentive Conditions (Reprinted with permission from Marsh, 1984b)

Condition	Expressiveness (%)	Content (%)	Interaction (%)
No External Incentive			
Clarity/Organization	11.3**	4.2**	1.6
Instructor Concern	12.9**	2.1	2.8*
Instructor Knowledge	12.8**	2.7*	1.9*
Instructor Enthusiasm	34.6**	1.9*	2.4*
Learning Stimulation	13.0**	9.6**	1.5
Total rating (across all items)	25.4**	5.1**	3.3**
Achievement test scores	9.4**	5.2**	1.3
Incentive After Lecture			
Clarity/Organization	2.0	6.0	1.3
Instructor Concern	20.5**	7.5**	1.9
Instructor Knowledge	25.1**	8.8**	2.3
Instructor Enthusiasm	30.9**	.1	3.3
Learning Stimulation	4.1*	2.9	.7
Total rating (across all items)	23.4**	7.0**	2.8
Achievement test scores	.3	13.0**	.4
Incentive Before Lecture			
Clarity/Organization	.3	11.5**	6.9*
Instructor Concern	.1	7.0*	6.2*
Instructor Knowledge	.3	6.2*	1.3
Instructor Enthusiasm	22.1**	4.0	6.6*
Learning Stimulation	.1	8.8*	8.1*
Total rating (across all items)	2.0	11.4**	6.8*
Achievement test scores	.5	26.5**	2.7
Across All Incentive Conditions			
Clarity/Organization	2.1**	5.0**	1.6*
Instructor Concern	7.2**	4.3**	1.0
Instructor Knowledge	6.4**	3.1**	.8
Instructor Enthusiasm	25.4**	1.2*	1.7**
Learning Stimulation	3.3**	4.9**	1.1
Total rating (across all items)	12.5**	5.2**	1.8*
Achievement test scores	1.7**	10.7**	.3

Note. Separate analyses of variance (ANOVAs) were performed for each of the five evaluation factors, the sum of the 18 rating items (Total rating), and the achievement test. First, separate two-way ANOVAs (Expressiveness × Content) were performed for each of the three incentive conditions, and then three-way ANOVA's (Incentive × Expressiveness × Content) were performed for all the data. The effect sizes were defined as $(SS_{effect}/SS_{total}) \times 100\%$.

* $p < .05$. ** $p < .01$.

BEST COPY AVAILABLE

Chapter 7: Utility of Student Ratings-- Improvement of Instruction.

The introduction of a broad institution-based, carefully planned program of students' evaluations of teaching effectiveness is likely to lead to the improvement of teaching because:

1) faculty will have to give serious consideration to their own teaching in order to evaluate the merits of the program;

2) the institution of a program which is supported by the administration will serve notice that teaching effectiveness is being taken more seriously by the administrative hierarchy.

3) the results of student ratings, as one indicator of effective teaching, will provide a basis for informed administrative decisions and thereby increase the likelihood that quality teaching will be recognized and rewarded, and that good teachers will be kept.

4) the social reinforcement of getting favorable ratings will provide added incentive for the improvement of teaching, even for tenured faculty.

5) faculty report that the feedback from students' evaluations is useful to their own efforts for the improvement of their teaching.

****None of these observations, however, provides an empirical demonstration of improvement of teaching effectiveness resulting from students' evaluations.

Feedback Studies.

In most studies of the effects of feedback from students' evaluations:

1) classes are randomly assigned to experimental or control groups;

2) students' evaluations are collected near the middle of the term;

3) at least the ratings from one or more groups are returned to instructors as quickly as possible;

4) the various groups are compared at the end of the term on a second administration of student ratings and as well as other variables.

In FEEDBACK studies using SEEQ in multiple sections of the same course:

Study 1. Results from an abbreviated form of the survey were simply returned to faculty, and the impact of the feedback was positive, but very modest.

Study 2. Here I actually met with instructors in the feedback group to discuss the evaluations and possible strategies for improvement. In this study students in the feedback group subsequently performed better on a standardized final examination, rated teaching effectiveness more favorably at the end of the course, and experienced more favorable affective outcomes at the end of the course (i.e., feelings of course mastery, and plans to pursue and/or apply the subject).

***These two studies suggest that feedback, coupled with a candid discussion with an external consultant, can be an effective intervention for the improvement of teaching effectiveness.

Remaining Issues

Several issues still remain for FEEDBACK research.

1) How much of the observed effect is due to consultation that does not depend on feedback from student ratings?

2) Nearly all of the feedback studies were based on midterm feedback from midterm ratings. This limitation, perhaps, weakens the likely effects in that many instructional characteristics cannot be easily altered in the second half of the course. This approach also requires further study of the generality of this approach to the effects of end-of-term ratings in one term to subsequent teaching that is more typical.

3) reward structure is an important variable which has not been examined in this feedback research. Even though faculty may be intrinsically motivated to improve their teaching effectiveness, potentially valuable feedback will be much less useful if there is no extrinsic motivation for faculty to improve. To the extent that salary, promotion, and prestige are based almost exclusively on research productivity, the usefulness of student ratings as feedback for the improvement of teaching may be limited.

4) There has been too little systematic research on the usefulness of students' evaluations for the other purposes for which they are intended: personnel decisions, student instructor/course selection, and research on teaching.

Table 9

F Values for Differences Between Students With Either Feedback or No-Feedback Instructors For End-of-Term Ratings, Final Exam Performance, and Affective Course Consequences (Reprinted with permission from Overall and Marsh, 1979; see original article for more details of the analysis).

Variable	Group: Feedback[a] *M*	Feedback[a] *SD*	No feedback[b] *M*	No feedback[b] *SD*	Difference	*F*(1, 601)
Rating components						
Concern	52.38	8.5	49.51	10.1	2.87	19.1**
Breadth	50.84	7.9	49.59	7.9	1.25	4.8*
Interaction	51.94	7.4	48.61	10.3	3.33	32.4**
Organization	49.88	9.4	50.88	9.5	−1.00	2.5
Learning/Value	50.77	9.9	48.22	10.7	2.55	11.7**
Exams/Grading	50.52	9.9	49.08	10.1	1.44	4.1*
Workload/Difficulty	51.13	8.8	51.51	8.8	−.38	.4
Overall Instructor	7.00	1.6	6.33	2.1	.67	26.4**
Overall Course	5.81	1.8	5.39	2.0	.42	5.4*
Instructional Improvement	5.97	1.5	5.49	1.5	.48	16.0**
Final exam performance	51.34	9.9	49.41	10.1	1.93	9.4**
Affective course consequences						
Programming competence achieved	5.80	2.0	5.42	2.3	.38	7.7**
Computer understanding gained	6.18	2.0	5.94	2.1	.24	3.6
Future computer use planned	4.00	2.8	3.49	2.7	.51	6.5*
Future computer application planned	5.05	2.6	4.67	2.6	.38	5.4*
Further related coursework planned	4.39	2.9	3.52	2.9	.87	11.1**

Note. Evaluation factors and final exam performance were standardised with $M = 50$ and $SD = 10$. Responses to summary rating items and affective course consequences varied along a scale ranging from 1 (very low) to 9 (very high). *F* test for the main effect of the feedback manipulation in the analysis is summarised in Table 3.

[a] For feedback group, $N = 295$ students in 12 sections.

[b] For no-feedback group, $N = 456$ students in 18 sections.

* $p < .05$. ** $p < .01$.

BEST COPY AVAILABLE

Chapter 8: The Use of Student Ratings In Different Countries: The Applicability Paradigm

Students' evaluations are collected in most North American Universities, but not in other parts of the world and not in secondary institutions. The Applicability Paradigm is designed to test the applicability of two rating instruments -- my SEEQ and Peter Frey's Endeavor -- to other countries. A representative sample of students is asked to:

a) select a "best" and a "worst" teacher,

b) rate each using SEEQ and Endeavor,

c) indicate inappropriate items, and

d) select the most important items

Analyses of the results included:

a) a discrimination of "best" and "worst" teachers

b) comparisons of "inappropriate" and "most important" items.

c) factor analyses of SEEQ and Endeavor responses

d) multitrait-multimethod analyses of relations between SEEQ and Endeavor scales

The applicability paradigm has been used in Spain, Papua New Guinea, New Zealand, Indonesia, and two different tertiary settings in Australia. In each study most items were judged to be appropriate and chosen by at least some as most important, and all but Workload/Difficulty items differentiated between good and poor teachers. There was a surprising consistency in the items chosen as most important and inappropriate across the studies. Factor analyses identified most of the factors the instruments were designed to measure. The MTMM analyses provided support for both the convergent and discriminant validity of the responses to the two instruments. The studies suggest that students in different countries do differentiate among different components of effective teaching in a way similar to North American students when responding to SEEQ and Endeavor.

Based on these studies, the Applicability Paradigm appears to provide a useful initial study in evaluating the applicability of students' evaluations of teaching effectiveness in a new setting.

Chapter 9: OVERVIEW, SUMMARY AND IMPLICATIONS

Research reviewed shows that student ratings are:

1) multidimensional;

2) reliable and stable;

3) primarily a function of the instructor who teaches a course rather than of the course that is taught;

4) relatively valid against a variety of indicators of effective teaching;

5) relatively unaffected by a variety of variables hypothesized as potential biases;

6) seen to be useful by faculty as feedback about their teaching, by students for course selection, and by administrators for use in personnel decisions.

However, the same findings also demonstrate that student ratings have some faults, and they are viewed with some skepticism by faculty as a basis for personnel decisions.

This level of uncertainty probably also exists for all personnel evaluations -- particularly among those being evaluated. Students' evaluations of teaching effectiveness are probably the most thoroughly studied form of personnel evaluation, and one of the best in terms of being supported by empirical research.

Alternative Indicators of Effective Teaching

Despite the generally supportive research findings, student ratings should be used cautiously. There should be other forms of systematic input about teaching effectiveness, particularly for personnel decisions.

Whereas there is good evidence to support the use of students' evaluations as one indicator of effective teaching, there are few other indicators of teaching effectiveness whose use is systematically supported by research findings.

Extensive lists of alternative indicators of effective teaching are proposed, but few are supported by systematic research, and none are as clearly supported as students' evaluations of teaching.